Science AAAS

Supplementary Materials for

Systematic mapping of functional enhancer-promoter connections with CRISPR interference

Charles P. Fulco[1,2], Mathias Munschauer[1], Rockwell Anyoha[1], Glen Munson[1], Sharon R. Grossman[1,3,4], Elizabeth M. Perez[1], Michael Kane[1], Brian Cleary[1,5], Eric S. Lander[1,2,4]*, Jesse M. Engreitz[1]*

*correspondence to: engreitz@broadinstitute.org and lander@broadinstitute.org

This PDF file includes:

Materials and Methods

Selection of targets for sgRNA library

To develop this CRISPRi screening approach (see Supplemental text), we focused on two genes — *MYC* and *GATA1* — that play critical roles in human development and disease and that are known to affect cellular proliferation in K562 cells (*26*). We determined by consulting a genome-wide catalog of gene essentiality in K562 cells (*26*) as well as Hi-C data in K562 cells (*6*) that *MYC* and *GATA1* are not located in close linear (500 Kb) or spatial proximity (within the same topological domain) to other genes expressed in K562 cells that strongly affect cell proliferation (**Fig. S1**). We also examined the potential effects of several noncoding RNAs in the *MYC* locus on cell proliferation, but determined that none are likely to contribute (see Supplemental text).

We designed an sgRNA library containing guides targeting several loci as well as internal controls, for a total of 98,599 sgRNAs (**Table S2**). We dedicated most of the sgRNAs in the library to studying the *MYC* locus, due to the apparent complexity of its regulatory architecture (*e.g.*, see **Fig. 3A**) (*27*) and its importance in many human cancers. To identify the elements that regulate *MYC*, we examined the 3-Mb topological domain and selected a ~666 Kb region that contained *MYC* itself, many elements with strong DHS and H3K27ac signal in K562 cells, and all intervening regions. We selected additional regions throughout the domain to cover other strong H3K27ac peaks downstream of *MYC* (including the regions surrounding e5-e7 that from Hi-C can be observed to form long-range loops to the *MYC* promoter), as well as additional regions upstream of *MYC* that are marked by active chromatin in other cell types but not in K562s (*e.g.*, see **Fig. 3A**). In each case, we included at least 5 kb of sequence surrounding the ENCODE "broadPeak" annotations. We note that performing similar experiments with larger libraries — for example including all possible sgRNAs in the the 3-Mb topological domain containing *MYC* — is possible and would require increasing the scale of the experiment (number of cells and reads) accordingly.

For *GATA1*, we tiled a 74 kb region containing the *GATA1* gene body as well as several putative enhancer elements nearby, including 17 kb annotated as "weak enhancer" and 19.4 kb annotated as "strong enhancer" by ENCODE ChromHMM (**Fig. 1B**). We note that we do not rule out the possibility that additional regulatory elements beyond this span may regulate *GATA1*.

We included several additional sets of sgRNAs as internal positive and negative controls for the screen. As negative controls, we included 4,082 scrambled-sequence sgRNAs, selected to include all 20- or 21-nucleotide sgRNAs from the previous genome-wide CRISPRi screening library designed by the Weissman lab (*10*), subject to the filters described below. We also included sgRNAs targeting the promoters of 600 protein-coding genes – including 535 that are expressed in K562 cells (fragments per kilobase per million >1) and 65 that are not expressed – as internal standards in the screen to compare to previous genome-wide screens assessing gene essentiality (*10*, *26*). We selected these genes to span the range of potential effects on cellular proliferation, including the 52 most essential genes reported previously (*26*).

Finally, because sgRNAs tiling across a noncoding region might be subject to different biases than scrambled-sequence sgRNAs (*e.g.*, due to specific sequence motifs, repetitive regions, or general toxic effects of targeting KRAB-dCas9 to chromatin), we selected additional negative control regions that are not close to genes known to be strongly essential but nonetheless do have putative regulatory elements marked by DHS and H3K27ac. We used these negative control regions (85 kb total) to estimate an empirical false discovery rate for elements in the *GATA1* and *MYC* loci (see below).

sgRNA design for tiling noncoding sequences

To design sgRNAs for tiling across noncoding sequences, we generated a list of all possible targeting sites with an NGG PAM. We calculated a specificity score based on potential off-target sites using a previously described algorithm (http://crispr.mit.edu, (*28*)), and removed guides with specificity scores <20. We note that this means that certain noncoding regions, including regions containing repetitive elements, are not tested by this screen. For cloning sgRNAs into sgOpti, we added a "G" base to the beginning of the 20-nucleotide sequence if the first base was not already a "G". We note that we applied additional filters to the sgRNAs considered during analysis of the screen (see below).

sgRNA design for targeting promoters

Because CRISPRi has a ~200-bp window of efficacy surrounding the TSS (Supplemental text) (*29*), we used capped analysis of gene expression (CAGE) data from K562 cells (*30*) to precisely define TSS locations (10-bp resolution) and designed sgRNAs targeting the regions immediately proximal to this site. In cases where genes showed multiple TSSs (as judged by the second-strongest TSS having >20% of the CAGE signal of the stronger TSS), we designed sgRNAs against both of these TSSs. To design sgRNAs targeting these sites, we used an algorithm based on a previous approach (*10*). We first generated all possible guides of length 18-24 where the first position in the genome corresponds to a "G", filtering out those with potential for off-target effects based on their specificity score. We defined prioritized windows around the TSS corresponding to (-30 to +45 bp), (-30 to +95 bp), and (-200 to +200 bp). We selected sgRNAs from these regions in order until we obtained 20 sgRNAs per promoter. For each window, we chose as many sgRNAs as possible that were spaced at least 5 bp apart, and then moved to the next priority window.

Tissue Culture

We maintained K562 (ATCC) cells a density between 100K and 1M per mL in RPMI-1640 (Thermo Fisher Scientific, Waltham, MA) with 10% heat-inactivated FBS (HIFBS, (Thermo Fisher Scientific), 2mM L-glutamine, and 100 units/ml streptomycin and 100 mg/ml penicillin. We maintained HEK293Ts between 20 and 80% confluence in DMEM with 1 mM Sodium Pyruvate, 25mM Glucose (Thermo Fisher Scientific) and 10% HIFBS unless otherwise noted.

Constructs for CRISPRi

We expressed sgRNAs from sgOpti, a modification of pLenti-sgRNA (Addgene #71409) with the sgRNA scaffold replaced with the sgRNA-(F+E)-combined optimized scaffold previously described (*31*). We generated constructs expressing inducible KRAB-dCas9 by replacing the SFFV promoter with a TRE3G promoter and the P2A-mCherry cassette with an IRES-GFP or IRES-BFP cassette in pHR-SFFV-KRAB-dCas9-P2A-mCherry (Addgene #60954) (*10*).

CRISPRi line generation

We generated the inducible CRISPRi cells lines by (i) transducing K562 cells with a construct expressing rtTA linked by IRES to a neomycin resistance cassette expressed from an EF1α promoter (ClonTech, Mountain View, CA) and selecting with 200 µg/mL G418 (Thermo Fisher), then (ii) transducing these rtTA-expressing K562 cells with one of the KRAB-dCas9 constructs described in the section above. We selected for cells expressing GFP or BFP by fluorescence activated cell sorting (FACS).

sgRNA library cloning

We synthesized an oligo pool corresponding to the sgRNA library with PCR tags (purchased from CustomArray, Bothell, WA, **Table S2**). We amplified the pool by PCR with primers sgRNA Library Fwd/Rev to add homology arms for Gibson assembly (**Table S3**), and purified the product with an equal volume (1×) AMPure XP SPRI beads (Beckman Coulter, Danvers, MA). We prepared the vector backbone by digesting sgOpti with BsmBI (New England Biolabs (NEB), Ipswich, MA) followed by purification with 0.75× AMPure XP SPRI. We assembled 70 ng amplified library into 500 ng digested vector in a 50 µL Gibson reaction (NEB), cleaned these by 0.75× AMPure XP SPRI, eluted in 15 µL H_2O and electroporated the entire volume into Endura competent cells (Lucigen, Middleton, WI). We expanded the cells in liquid culture for 18 hours at 30 °C and purified the pooled library plasmid with the Endotoxin-Free Plasmid Maxiprep Kit (Qiagen, Hilden, Germany).

Lentivirus production

We plated 700,000 HEK293T cells on 6-well plates (Corning, Corning, NY) and 24 hours later transfected with 1 µg dVPR, 300 ng VSVG, and 1.2 µg transfer plasmid using XtremeGene9 (Roche Diagnostics, Indianapolis, IN). For pooled s, the cell number and plasmid mass were scaled proportionally to 14 million cells on a 15 cm plate (Corning). 16 hours post-transfection we changed media to DMEM with 20% HIFBS. At 48 hours post-transfection, we harvested viral supernatants and filtered them through a 0.45 µM syringe filter before use.

Pooled CRISPRi screens for essentiality

We transduced K562 harboring a doxycycline-inducible KRAB-dCas9 at an multiplicity of infection (MOI) of 0.3 at a coverage of 1,000 transduced cells per sgRNA as previously described (*26*). Starting 36 hours after transduction, we selected for successfully transduced cells with 1 µg/mL puromycin for 72 hours and collected 150 million cells as a reference sample. After maintaining cells at 1,000× coverage in 0.2 µg/mL puromycin and 0.5 µg/mL doxycycline for 14 population doublings, we collected 150 million cells of the final cell population. We extracted genomic DNA from both the reference and final cell populations using the QIAamp DNA Blood Maxi kit (Qiagen) according to the manufacturer's instructions. We amplified sgRNAs integrations from 900 µg genomic DNA by PCR with indexed sgRNA sequencing library primers containing Illumina adaptors (**Table S3**) and sequenced them on a HiSeq 2500 using custom Illumina sequencing and index primers (**Table S3**) to an average depth of >350 reads per sgRNA. We used Bowtie (*32*) to align the resulting sequences to the sgRNA library allowing perfect matches only.

Analysis of sgRNA depletion in proliferation-based screen

To evaluate the potential of off-target sgRNA-mediated toxicity to affect cellular proliferation, we inspected the depletion of the set of sgRNAs in the tiled negative control regions (where we expect no on-target sgRNA depletion) and noted that the frequency of sgRNAs more than 2-fold depleted across the screen is higher (2-proportion Z-test $p<0.0001$) in sgRNAs with specificity scores below 50 (9%) than those with a score of 50 or above (5%). We considered only the sgRNAs with specificity scores >50 in the subsequent analysis. We also ignored sgRNAs with more than 10 “G” bases in the targeting sequence, which also lead to an increased frequency of off-target toxicity based on analysis of the negative control sgRNAs. These filters retain >90% of sgRNAs. To ensure robust calculation of sgRNA scores, we examined only sgRNAs with at least 50 raw reads in the initial timepoints for both replicates (retains 98% of sgRNAs). We assessed the depletion of the remaining sgRNAs as described below.

CRISPRi score

The “CRISPRi score” represents the $-\log_2$ depletion between the beginning and end of the proliferation screen (14 doublings). We calculated the CRISPRi score for each of two replicates and report the mean of these scores as the CRISPRi score for each sgRNA. To identify significant regions by integrating information from multiple sgRNAs, we used a sliding window approach, averaging the mean CRISPRi score across N consecutive guides. To choose N, we compared the correlation of the window CRISPRi scores between the two replicates as a function of N (**Fig. S2A**). We found that using $N = 20$ yielded a Pearson’s correlation of 0.80 between the two replicates (**Fig. S2B**). As the sgRNAs were spaced on average every ~16 bp (**Fig. S2C**), windows of 20 consecutive sgRNAs spanned on average 314 bp (median = 237 bp, **Fig. S2D**). We note that this resolution is on the same order as the size of scoring regions in our CRISPRi screen (hundreds of bp), indicating that choosing a smaller window size would not necessarily increase the resolution of the approach. Because some regions are covered sparsely due to repetitive sequence, we considered windows only if they contained 20 guides within a span of 1000 bp (**Fig. S2D**). We note that the enhancers we identify (e-GATA1, e-HDAC6, e1-e7) are robust to the precise choice of window size.

To identify significant windows, we required first that the CRISPRi score for the window had an irreproducible discovery rate < 0.05 (*33*) when comparing the two replicate screens. Second, we tested whether the mean of the sgRNAs in each window deviated significantly from the mean of the negative controls, using sgRNA CRISPRi scores averaged across duplicate screens. Specifically, we calculated a T-test statistic by comparing the CRISPRi scores of the 20 sgRNAs with those of the scrambled-sequence, negative control sgRNAs. We assessed the empirical false discovery rate (FDR) of windows in the GATA1 and MYC loci by comparing these T statistics to those generated from sliding windows across three negative control regions that are located far from known essential genes expressed in K562 (see Selection of targets for sgRNA library), and selected a threshold based on a FDR of 0.05. This threshold corresponded to a

Benjamini-Hochberg-corrected T-test *p*-value of 0.032. We considered significant elements with an absolute effect size of >25%.

The final reported CRISPRi scores for 20-sgRNA windows in figures and **Table S2** represent the average of the two replicate screens normalized to the average of the scrambled-sequence negative-control sgRNAs.

Sources for epigenomics data

We downloaded data generated by the ENCODE Project Consortium (*4*) in K562 cells corresponding to DNase I hypersensitivity sequencing (DHS-seq); H3K27ac, GATA1, and CTCF chromatin immunoprecipitation sequencing (ChIP-seq); the chromatin state hidden Markov model (ChromHMM); and RNA Pol II ChIA-PET (*3*). To examine transcription factor occupancy at various enhancers, we downloaded the genome-wide binding sites of 100 transcription factors based on ChIP-Seq in K562 cells (wgEncodeRegTfbsClustered track from UCSC Genome Browser). We obtained sequence conservation from the UCSC Genome Browser corresponding to the phastCons 100-mammal multiple alignment (*34*). CTCF motifs were identified using FIMO (*35*) to search for the "V_CTCF_01" and "V_CTCF_02" position weight matrices from TRANSFAC (*36*). We obtained *in situ* Hi-C data for multiple cell types and used 5-Kb resolution KL-normalized observed matrix for all plots and analyses (*6*).

Cloning individual sgRNAs

For each of the selected enhancers (e-GATA1, e-HDAC6, e1-e7), and promoters (*GATA1* and *MYC*) that scored in the screen, we selected 2 non-overlapping sgRNAs with a preference for sgRNAs with high specificity and CRISPRi scores and sgRNAs that overlap the peak of DNase hypersensitivity. For regions that did not score (NS1, *HDAC6* promoter), we selected sgRNAs based on the same criteria, although because these sgRNAs were not high scoring, we also preferred guides predicted to have high efficacy (*37*). As negative controls, we selected 5 sgRNAs from the set without genomic targets. We cloned these sgRNAs as previously described (*38*) into sgOpti.

Generating sgRNA-expressing stable cell lines

We generated stable cell lines expressing single sgRNAs by lentiviral transduction in 8 μg/ml polybrene by centrifugation at 1400 x *g* for 45 minutes with one million cells per well in 24 well plates. After 24 hours, we selected for transduction with 1 μg/ml puromycin (Gibco) for 72 hours then maintained cells in 0.2 μg/ml puromycin. For each sgRNA, we generated three independent polyclonal cell populations through triplicate infections.

Single sgRNA knockdown

We plated sgRNA-expressing stable cell lines at 200,000 cells/ml in 0.5 μg/ml doxycycline and harvested cells 24 hours later by lysing in Buffer RLT (Qiagen).

RNA extraction and quantitative RT-PCR

We extracted RNA from 20,000-50,000 cells per experiment in Buffer RLT (Qiagen) using Dynabeads MyOne Silane beads (Thermo Fisher), treated samples with TURBO DNase (Thermo Fisher), and cleaned again with Dynabeads MyOne Silane beads. We used AffinityScript reverse transcriptase (Agilent Technologies, Lexington, MA) and random nonamer primers to convert RNA to cDNA. We performed qPCR using SYBR Green I Master Mix (Roche) and calculated differences using the ΔΔCT method versus GAPDH (see **Table S3** for primer sequences).

To achieve power to detect small effects in gene expression, we performed 3 technical qPCR replicates (from the same cDNA) and took the median value for further analysis. We also included many biological replicates. Specifically, we derived 3 independent lines for each sgRNA and assayed each once as a biological replicate in GATA1 locus experiments (for a total of 3 replicates) and 4 times for experiments in the MYC locus (for a total of 12 biological replicates)

RNA sequencing and analysis

To examine the transcriptional changes resulting from inhibition of a GATA1 enhancer, we performed RNA-sequencing on cell lines expressing individual sgRNAs targeting the GATA1 TSS (2 different sgRNAs), e-HDAC6 (2 different sgRNAs), and non-targeting, negative controls (4 different sgRNAs). We generated RNA sequencing libraries from 3 biological replicates for each sgRNA and processed the data as previously described (*39*). We identified differentially expressed genes ($q < 0.05$, fold-change > 2) with DESeq2 (version 1.6.3) (*40*) and found a significant overlap in the sets of differentially expressed genes between GATA1 TSS and e-HDAC6 targeting sgRNAs (**Fig. S4B**), suggesting that e-HDAC6 leads to downstream transcriptional changes consistent with direct regulation of GATA1.

Single sgRNA competitive growth assays

For competition experiments we pooled the indicated K562 cells expressing an individual sgRNA and KRAB-dCas9-IRES-BFP with K562s expressing either GFP or RFP (control cells) in 0.5 μg/mL doxycycline. We measured the fractions of CRISPRi and control cells by flow cytometry after 24 hours and again after 7 additional days. We performed each experiment in six replicates including competitions against both the GFP- and RFP-expressing control lines. We quantified the growth phenotype gamma as previously described (*10*).

Luciferase reporter assays for enhancer activity on a plasmid

To test the functions of each putative regulatory element in a classic reporter-based enhancer assay, we created a reporter plasmid derived from pGL4.23 (Promega, Madison, WI) where firefly luciferase is expressed from a 180-bp fragment of the *MYC* promoter (hg19 coordinates: chr8:128748316-128748495). We designed an insertion site ~2 kb upstream of the *MYC* promoter for inserting each candidate enhancer sequence, and we flanked this region with polyadenylation signals in either direction to avoid measuring luciferase activity driven from transcripts initiating from the enhancer elements themselves. Primers for each element tested are listed in **Table S3**. The negative control sequence corresponded to a kanamycin resistance cassette.

For each construct, we transfected 500,000 K562 cells using the Lonza (Cologne, Germany) Amaxa 96-well Shuttle according to the manufacturer's instructions for this cell type (except transfecting all 500,000 cells in a single well) with 250 ng of reporter plasmid plus 250 ng of a plasmid expressing *Renilla* luciferase. We harvested cells 48 hours after transfection by spinning once, washing with PBS, and resuspending in 40 µl Passive Lysis Buffer (Promega). We performed the Dual-Luciferase Reporter Assay according to the manufacturer's protocol (Promega). Barplots report firefly luciferase activity normalized to *Renilla* luciferase activity and to the negative control construct for 3 replicate transfections.

Chromatin immunoprecipitation for H3K27ac

We performed ChIP for H3K27ac as previously described, with modifications (*41*). We grew K562 cells expressing individual sgRNAs targeting *MYC* enhancers or negative controls in the presence of doxycycline for 48 hours. We harvested cells, washed once in cold PBS, and crosslinked with 1% formaldehyde in PBS for 10 minutes at 37 °C followed by quenching with glycine for 5 minutes at 37 °C. We washed cells twice in ice cold PBS with 1× protease inhibitor (Roche). We flash froze the pellets and stored at -80°C until sonication, at which time we thawed the pellets on ice and lysed cells in ChIP Lysis Buffer (1% SDS, 10 mM EDTA, 50 mM Tris-HCl pH 8.0) on ice for 10 minutes. We sonicated batches of 3 million cells in 100 µL using a Q800R2 Sonicator (QSonica, Newtown, CT) at 50% amplitude, 30 s on / 30 s off, for 7.5 minutes to obtain fragment sizes between 150 and 700 bp.

We diluted 100 µL lysate from 1 millions cells in 660 µL ChIP Dilution Buffer (0.01% SDS, 1.1% Triton X-100, 1.12 mM EDTA, 16.7 mM Tris-HCl pH 8.0), and saved an aliquot for whole-cell extract. For immunoprecipitation of H3K27ac (using antibody 39685 from Active Motif, Carlsbad, CA), we incubated 5 µl of antibody with Protein A/G beads (Thermo Fisher) in Blocking Buffer (500 mM Tween-20, 500 mM BSA in 1x PBS) for 2 hours at 4 °C. We then washed the beads once in Blocking Buffer, resuspended the beads in 55 µL Blocking Buffer, and added it to the DNA samples. We incubated the antibody-bead-lysate mixture overnight at 4°C rotating end over end. Next day, we washed the samples as follows: four times with 200 µL of RIPA Buffer (0.1% Na-deoxycholate, 0.1% SDS, 1% Triton X-100, 100 mM NaCl, 1 mM EDTA, 10 mM Tris-HCl pH 8.0), twice with 100 uL RIPA High Salt Buffer (0.1% Na-deoxycholate, 0.1% SDS, 1% Triton X-100, 500 mM NaCl, 1 mM EDTA, 10 mM Tris-HCl pH 8.0), twice with LiCl Wash Buffer (250 mM LiCl, 0.5% NP-40, 0.5% Na-deoxycholate, 1 mM EDTA, 10 mM Tris-HCl pH 8.0), and twice with 1× TE. Following the washes, we resuspended beads in Elution Buffer (10 mM Tris-HCl pH 8.0, 5 mM EDTA, 300 mM NaCl, 0.1% SDS) and incubated the resuspended beads at 65 °C for 10 minutes. Following this first brief reverse crosslinking step, we added 5 µL RNase Cocktail (Thermo Fisher) and incubated at 37 °C for 30 minutes, and then added 5 µl Proteinase K (NEB) and incubated at 65 °C for 2 hours. Samples were cooled on ice. DNA was extracted using Agencourt XP (SPRI) beads (Beckman Coulter) at 2× sample volume, followed by elution in 10 mM Tris-HCl pH 8.0. We performed quantitative PCR using Roche 2× SYBR Green Master Mix on a Roche LightCycler 480. We calculated enrichment compared to 5 positive control primers designed against H3K27ac peaks outside of the *MYC* region. Primer sequences are listed in Table S3.

siRNA-mediated knockdown of MYC, GATA1, and PVT1

We transfected 200,000 cells with 10 nM siRNAs obtained from GE Dharmacon (Lafayette, CO, **Table S3**) in quadruplicate using the Neon transfection system (Thermo

Fisher, settings: 1,450 V, 10 ms width, 3 pulses). We harvested cells in Buffer RLT (Qiagen) 24 hours after knockdown and estimated target gene expression relative to cells transfected with non-targeting siRNAs by quantitative PCR as described above. For competition experiments we transfected fluorescently labeled cells (GFP or RFP) with indicated siRNAs at 10 nM following the described procedure. We pooled cells such that cells transfected with siRNAs targeting PVT1, MYC or GATA1 were matched with differently labeled cells transfected with non-targeting control siRNAs. We measured the GFP and RFP fractions immediately following transfection and again after 4 days by flow cytometry. Each experiment was carried out in quadruplicates and included a label-swap experiment.

Strategy for genetic deletions of enhancers in the *MYC* locus

To test the effects of enhancers on *MYC* expression through genetic manipulations, one straightforward experiment would be to use CRISPR/Cas9 to generate clonal cell lines containing homozygous knockouts of each putative enhancer and measure the effects on *MYC* using the qPCR assays described above. However, there are several reasons why this experiment is not ideal in our system. First, we observe significant biological variation in MYC expression between clonal cell lines. Second, MYC affects cellular proliferation and thus cells lacking one of these enhancers may be outcompeted. Finally, K562 cells are triploid, making it difficult to obtain cell lines where an enhancer is removed on all 3 alleles.

Accordingly, we developed an alternative strategy (**Fig. S7**). We used CRISPR/Cas9 to generate clonal cell lines carrying *heterozygous* genetic deletions (on 1 or 2 of the 3 homologous chromosomes) and compared the expression of *MYC* on the modified and unmodified homologous chromosomes in the same cells. We expect that if the enhancer in fact regulates *MYC*, *MYC* expression from the modified allele should be reduced compared to the wild-type allele. This approach is identical in concept to classical *cis-trans* tests. This allele-specific approach can demonstrate that regulation of *MYC* is a direct, *cis* effect of the enhancer rather than an indirect effect (for example, due to the enhancer regulating another gene that in turn regulates *MYC*).

To implement this strategy, we first generated a cell line containing polymorphic sites on each allele of *MYC*. Because K562 cells do not contain polymorphisms in the *MYC* transcript, we knocked in polymorphic tags using CRISPR/Cas9 and homologous recombination. We first chose a targeting site in a *MYC* intron in a region that did not show sequence conservation across mammals. We reasoned that editing such a site would not likely affect the regulation of *MYC*. We designed an sgRNA targeting this site (**Table S3**) as well as a ssDNA oligo to use as a donor for homologous recombination (**Fig. S7A**). This oligo contained four random nucleotides (NNNN), allowing us to generate cell lines containing unique polymorphic on each of the 3 alleles. We co-transfected these sgRNAs, Cas9, and the donor oligo in K562 cells, isolated clonal cell lines through serial dilution, and genotyped this intronic site by PCR and sequencing (for genotyping primers see **Table S3**). We identified a clonal cell line containing 3 distinct variants (CTAA, CCCG, and ATCG) in the targeted location. We expanded this cell line (K562-MYC-Tag) and used it for the second round of transfections.

To delete *MYC* enhancers, we designed sets of 4 sgRNAs flanking each element, with 2 sgRNAs on each side. These sgRNAs were designed to delete ~1 kb regions containing the DHS site in the middle of the element. For e3 and e4, we designed the sgRNAs to cut outside of the exons and splice sites of PVT1. We co-transfected the K562-MYC-Tag cell line with Cas9 and sets of 4 sgRNAs, generated clonal cell lines through serial dilution, and genotyped each clone (**Fig. S7B**). We expanded clones containing deletions on 1 or 2 of the 3 alleles.

For each deletion clone and for 26 wild-type control clones, we use a droplet digital PCR (ddPCR) hydrolysis assay to measure the allele-specific expression of *MYC* and *PVT1*. We used this data, in combination with the genotyping amplicon sequencing, to infer partial phasing of the alleles relative to the polymorphic tags in the *MYC* intron (**Fig. S7C**). We performed these experiments for e2, e3, and e4 because these loci had SNPs that allowed us to determine which allele was deleted (see below). We compared the allele-specific expression between wild-type and deletion clones to determine how deleting *MYC* enhancers affected MYC expression (**Fig. S7D,E**).

Additional technical details for each of these steps are included below.

CRISPR/Cas9 transfections and clonal cell line selection

To delete specific sequences, we co-transfected 600 ng of Cas9-expressing plasmids (“PX330-NoGuide”), 300 ng of a pool of sgRNA-expressing plasmids (“pZB-Sg3”), and 600 ng of a plasmid expressing EGFP and a puromycin selectable marker from a CAG promoter (pS-pp7-GFPiP). To create PX330-NoGuide, we modified PX330 (gift from Feng Zhang, Addgene plasmid #44230) (*42*) to remove the sgRNA expression cassette. To generate pZB-Sg3, we cloned a human U6 promoter and optimized sgRNA scaffold sequence (*31*) into a minimal vector with an ampicillin-selectable marker and a ColE1 replication origin. We transfected batches of 250,000 human cancer cells using the Neon Transfection System (Invitrogen), using 3 pulses of 10 milliseconds at 1450 V and plated them into a 96-well plate in 200 µl media. As an internal control for each set of transfections, we performed a transfection using a pool of 4 sgRNAs with no predicted target sites in the human genome. To knock in polymorphic tags into the *MYC* locus, we included 200 ng of ssDNA oligo in the transfection (see **Table S3** for sequences).

We verified efficient transfection by examining GFP expression after 24 hours. To select for transfected cells, we replaced the media 24 hours after transfection with 200 µl media + 4 µg/ml puromycin. One day later, we split the cells into a 6-well plate with 2 ml of 4 µg/ml puromycin. One day later, we replaced the media with 2 ml of media with no puromycin. We allowed cells to grow for 7-8 days, replacing the media every 2-3 days. Once the cells could be reliably counted, we plated 8 96-well round-bottom plates at a dilution of 0.4 cells/well. We grew these plates in 200 ul of 20% FBS media, doing partial media changes every 3-4 days, for 12-16 days. Clonal cell lines were split into multiple copies and grown for 2-14 days before harvesting for biological replicates. We harvested cells for DNA and RNA extraction by removing most of the media and adding 3.5× volume Buffer RLT (Qiagen).

Genotyping deletion clones by PCR and sequencing

To genotype K562 clones, we isolated genomic DNA using Silane beads.

For genotyping MYC-Tag insertion clones (**Fig. S7A**), we performed PCR using primers (**Table S3**) surrounding the site followed by a second round of PCR to add a different barcode to each sample and sequenced the amplicons on an Illumina MiSeq (Illumina, San Diego, CA).

For genotyping deletion clones, we performed a first round of PCR using primers spanning the deleted region (**Fig. S7B**) and examined this PCR product using gel electrophoresis. Both wild-type and deletion-sized bands were visible and were used to prioritize clones for further analysis. We next performed a second nested PCR on this product to add sequencing tags and clone-specific barcodes for high-throughput sequencing (primers in **Table S3**). We sequenced these products to span the deletion junction; the number of unique amplicons in each clone was used to determine the number of deleted alleles. (This number is technically a lower bound, because in rare cases multiple alleles could be deleted and repaired in the same fashion). Finally, we counter-screened deletion clones for inversions, which can occur when Cas9-mediated cuts occur on both sides of the region, but the cuts are repaired with an inversion of the intervening sequence. We sought to eliminate clones that showed evidence of inversions, which could confound later analysis. For e2, we used primers spanning one side of the intended junction (**Table S3)** and eliminated clones that showed evidence of an amplicon corresponding to an inverted sequence. For e3 and e4, we were unable to obtain satisfactory PCR primers and so used a restriction digest approach that could distinguish whether the internal sequence was inverted or not. For e3, we digested PCR amplicons with AvrII and PsiI; for e4, we digested with NdeI and BglII (all enzymes from NEB).

Measuring allele-specific MYC and PVT1 expression in deletion clones

We designed and validated ddPCR assays to measure the allele-specific expression of MYC and PVT1. We first cloned the polymorphic regions of MYC and PVT1 from K562-MYC-Tag using the ddPCR-MYCIntron Fwd/Rev and ddPCR-PVT1 Fwd/Rev PCR primers (**Table S3**) to generate separate plasmid vectors containing each allele of each amplicon. We generated synthetic standard curves by mixing these vectors in specified ratios: 100:0, 90:10, 50:50, 10:90, and 0:100. Each standard curve was generated and quantified in duplicate to confirm that the assays were specific and quantitative.

To perform the ddPCR assay, each 20µl reaction contained 1X ddPCR Supermix for Probes - no dUTP (BioRad, Hercules, CA), 450 nM each of forward and reverse primer, and 500 nM probe. To measure the relative expression of the 3 MYC alleles (**Fig. S7C**), we used MYCIntron Fwd and Rev (**Table S3**) along with a FAM-conjugated CTAA or ATCG probe and a HEX-conjugated CCCG probe in two separate assays, then merged the results by comparing to the constant CCCG probe. To measure the relative expression of the 2 PVT1 polymorphisms (**Fig. S7C**), we used PVT1 Fwd and Rev and probes against T and C alleles in a single assay (**Table S3**). Probes were purchased as Custom ZEN Double-Quenched Probes (IDT). Following droplet generation on a QX200 droplet generator (BioRad), we performed 40 cycles of PCR with a 10 minute 55°C combined and melting extension step. We counted droplets using the QX200 Droplet Reader (BioRad) and determined allele specific expression by the ratio of FAM and HEX positive droplets.

To measure the allele-specific expression of each deletion clone, we generated cDNA from cells as described above and performed ddPCR using 1000 cell-equivalents of cDNA for *MYC* and 100 for *PVT1*. We measured each clone using 2 or 3 technical replicates and averaged the ratios between these measurements for further analysis.

Analysis of allele-specific expression data for deletion clones

To analyze the allele-specific ddPCR data for the deletion clones, we first inferred the phasing of the deletions relative to the polymorphic tags in *MYC*. We identified known polymorphisms near the deleted enhancers that would allow us to phase the deletions by examining DNA sequencing experiments from multiple types of ENCODE experiments (*e.g.*, ChIP-Seq, DHS sequencing). We identified rs67423398 (C/T/T in triploid K562 cells) just outside of the sgRNAs designed at e2 (**Fig. S7B**), allowing us to directly genotype the deletion bands by amplicon sequencing. For e3 and e4, there were no SNPs in the vicinity of the deletions themselves, but, because each acts as a promoter for *PVT1*, we were able to use a SNP in a downstream *PVT1* exon (rs11604, T/C/C in K562 cells) that allowed us to determine the allele of the deletions by examining which allele of PVT1 RNA was decreased (**Fig. S7C**). Accordingly, for each e2 clone we performed amplicon sequencing as described in the previous section and determined on which allele(s) the deletion occurred, and for each e3 and e4 clone we performed ddPCR to read out the allele-specific RNA expression of *PVT1*. This allowed us to determine whether the deletion occurred on the unique allele (C for rs67423398 or T for rs11604, C-T) or the ambiguous allele (T for rs67423398 or C for rs11604).

We next phased these polymorphisms based on the unique allele to the polymorphic tags in *MYC*. To do so, we first examined clones that carried deletions on the unique allele and examined their allele-specific expression of *MYC*. For e2, for example, we had 6 independent clones carrying such deletions, and these showed a consistent decrease in *MYC* expression on the CTAA allele (*e.g.*, **Fig. S7D**). We similarly linked the PVT1 unique allele to CTAA (**Fig. S7C**). By this strategy, we were able to phase some of the deletions to a unique *MYC* polymorphism (CTAA-C-T allele, **Fig. S7C**), and the remaining deletions to one of the other two alleles.

For each clone, we then calculated the change in expression of each *MYC* allele relative to 26 wild-type control clones. We first calculated the average expression of each allele in the control clones, which was approximately balanced (31% CTAA, 39% ATCG, 30% CCCG, **Fig. S7D**). For each clone, we compared the allelic expression fraction to the control clones to determine a fold-change for each allele. We then normalized these fold-changes to maximum of the 3 alleles, assuming that this represents a wild-type allele (*e.g.*, **Fig. S7D**, right), and termed this the “normalized allele expression”. We performed a similar computation on each wild-type clone. Finally, we compared the normalized allele expression between wild-type and deletion clones. For the unique allele (CTAA-C-T), we directly used the *MYC* normalized allele expression. For the remaining alleles (ATCG-T-C and CCCG-T-C), we chose the one of the two alleles with the lowest normalized allele expression, assuming that this was the deletion allele, and similarly generated a distribution of control values by performing a similar procedure on wild-type clones. We combined these comparisons across alleles and compared deletion to control clones using a Wilcoxon rank sum test (**Fig. S7E**).

Comparison to previous enhancer-promoter predictions

Given our functional mapping of enhancers that regulate *MYC*, we compared our list of true *MYC* enhancers to existing methods for predicting or inferring enhancer-promoter connections. We found that none of these strategies specifically identified more than 2 of the 7 *MYC* enhancers and correctly distinguished the 2 *GATA1* enhancers from neighboring elements that do not affect *GATA1* expression. We describe each of these approaches below.

1. One commonly used strategy for connecting enhancers with target promoters is to assign an enhancer to its nearest gene. It is clear that this does not accurately capture the complexity of enhancer-promoter connections (*8*), but lacking clear alternatives this approach is frequently used to assess which gene an enhancer might regulate. For *GATA1*, this approach does not accurately capture how both e-GATA1 and e-HDAC6, which are closest to *GATA1* and *HDAC6*, respectively, in fact regulate both genes. For *MYC*, e1-e4 would be assigned as regulators of *PVT1*, while e5-e7 would be assigned to the *CCDC26* pseudogene.

Several methods for predicting enhancer-promoter connections are based on correlations in chromatin state across cell types.

2. One such method is based on correlation in histone modification profiles between candidate enhancer-promoter pairs within 125 kb across nine cell types, including K562 cells (*43*). Because of this distance restriction, this method does not make any predictions for *MYC*. For *GATA1*, this strategy misses both e-GATA1 and e-HDAC6, and makes dozens of incorrect predictions.
3. A second method based solely on correlation predicts enhancer-promoter pairs using correlation in DHS for all candidate pairs within 500 kb of one another across 125 cell types, including K562 cells (*44*). For *GATA1*, this method correctly identifies both e-GATA1 and e-HDAC6 but also incorrectly assigns two additional distal enhancers in the regions tested in our screen. For *MYC*, this approach correctly identifies only one of the K562 enhancers (e4) and makes dozens of other predictions that do not overlap e1-e7. (The published catalog from this study does not report which cell type each prediction refers to, and thus some of these additional predicted enhancers may represent regions that regulate one of the target genes in another cell type.)
4. A third correlation-based method (PreSTIGE) predicts enhancer-promoter pairs by pairing cell-type-specific H3K4me1 signals with cell-type specific gene expression across 12 cell types, using a 100 kb distance plus a subset of CTCF sites to set domain boundaries (*45*). In the *GATA1* locus, PreSTIGE reports that 29 kb of the 74 kb covered by our screen is an enhancer for *GATA1*, including both e-GATA1 and e-HDAC6 but incorrectly reporting many kilobases of additional sequence. In the *MYC* locus, PreSTIGE predicts a single region to be an enhancer; this region does not correspond to any of the enhancers we identify.

In addition to methods based on correlations in chromatin state across cell types, a second category of approaches for inferring enhancer-promoter functional connections is based on measuring their physical interactions with methods based on chromosome conformation capture. Physical contacts between enhancers and promoters correlate with gene activation (*1*, *6*, *46*, *47*), and in a few cases increasing the frequency of enhancer-promoter contact has been shown to activate gene expression (*48*, *49*). However, long-distance chromatin loops can form without regulatory effects on gene expression (*e.g.*, when a promoter forms a loop with a region that is not an enhancer), and the abilities of various features of chromosome conformation data to predict functional interactions remains unclear (*47*). Accordingly, we examined several features previously noted to correlate with enhancer-promoter connections to determine if they might correctly identify enhancers in the *MYC* locus.

5. We first examined loops as defined by *in situ* Hi-C (*6*). In a Hi-C map of K562 cells at 5 kb resolution, five focal loops involving the *MYC* promoter were reported. Of the five, one corresponds to the long-range loop with e6/e7, one corresponds to NS1, and the other three correspond to CTCF-bound sites that do not overlap *MYC* enhancers. Thus, at the reported significance thresholds and with the available resolution, these calls do not correspond with the enhancers that regulate *MYC*. Nonetheless, Hi-C data shows that these sites frequently contact *MYC* (**Fig. 2A**), and higher resolution maps may allow identification of focal loops to these sites. Regardless of the specific loop calls, we find that incorporating this information into our heuristic helps to rank enhancers likely to regulate *MYC* (see main text).
6. RNA Pol II ChIA-PET has been proposed as a proximity interaction method that enriches for enhancer-promoter interactions (*3*). ChIA-PET in K562 cells (wgEncodeGisChiaPetK562Pol2InteractionsRep1) identifies many interactions between *MYC* and sites throughout the adjacent contact domain (**Fig. 2A**). Notably, these do include all 7 of the *MYC* enhancers in K562, but also include dozens of other sites with equal or higher interaction frequencies (**Fig. 2A**). Furthermore, ChIA-PET in K562 cells does not detect interactions between *GATA1/HDAC6* and either of their enhancers.
7. Various methods developed to predict enhancer-promoter interactions have been developed and trained based on interactions identified in chromosome conformation capture experiments. Consistent with the poor positive predictive value of chromosome conformation capture data as described above, methods trained on this data (*e.g.*, (*50*, *51*)) also do not correctly identify *MYC* or *GATA1* enhancers.

Together, these observations highlight the importance of direct functional mapping of regulatory elements. Furthermore, they underscore the opportunity for new models that integrate these two classes of approaches based on chromatin state and proximity interactions in the context of appropriate training data generated through CRISPRi-based mapping of regulatory elements.

Calculating predicted impact of *MYC* enhancers in K562 cells

To rank the relative importance of putative activating elements near *MYC* in K562 cells, we first created a list of putative regulatory elements in the locus. We downloaded DHS peak calls from ENCODE (narrowPeak files corresponded to both replicates in K562 cells), expanded these peaks by 500 bp, and merged overlapping peaks. For each of these merged peaks, we calculated normalized read count (reads per million, RPM; *not* normalized to length of the element) from H3K27ac and DHS measurements in K562 cells, and retained windows in the top 50% percentile with respect to H3K27ac signal, yielding 93 putative regulatory elements. For each element, we calculated the normalized contact frequency to the *MYC* promoter by consulting KL-normalized observed contact matrices at 5-kb resolution generated by *in situ* Hi-C (*6*). We calculated relative impact by the following formula: Predicted impact = log_2(H3K27ac RPM × DHS RPM × Hi-C contact × Hi-C contact), thereby weighting "activity" and "proximity" approximately equally. Each element was ranked according to this score. In **Fig. 2E**, peaks overlapping the MYC enhancers were colored red and plotted versus their CRISPRi score, defined by the maximum CRISPRi score in a window overlapping the element.

To compare the performance of this heuristic with simpler models, we calculated rankings based on H3K27ac ChIP-Seq RPM only, DHS RPM only, and Hi-C contacts only for the same set of 93 putative regulatory elements (**Fig. S8A**). We note that because these 93 elements were selected based on DHS and H3K27ac signal as described above, this may be an optimistic estimate of the value of each dataset alone.

Additional experimental data will be required to further refine this model and determine whether it is applicable to different gene loci.

Calculating enhancer ranks across cell types

To expand this approach across additional cell types, we downloaded DHS and H3K27ac ChIP-seq data for diverse cell lines and primary tissues from the Roadmap Epigenomics Project (*5*), ENCODE (*4*), and others (*52*, *53*). While these data are available across a wide range of cell types (235 samples total), proximity interactions maps are available in a very limited number of cell types. Accordingly, we explored to what extent the topological architecture of the *MYC* locus changes across 7 human cell types previously mapped using *in situ* Hi-C (*6*, *54*). We found that key features of the proximity contacts of the *MYC* promoter appeared consistent across cell types, including the long-range contacts to the edges of the topological domain as well as several distinct peaks within these domains (**Fig. S8C**). These cell-type invariant long-range loops typically corresponded to sites bound by CTCF across multiple cell types, consistent with previous reports (*6*). Beyond these long-range loops, the quantitative interactions of the *MYC* promoter did change somewhat across different cell types, with elevated contact frequency coinciding with the presence of strong H3K27ac occupancy in a given cell type. To capture the features consistent across cell types, we generated a generic proximity profile for the *MYC* locus by averaging the proximity interactions across these 7 cell types, normalizing the absolute magnitude of interactions in each cell type by the signal at the *MYC* promoter itself. This generic profile accurately captured the cell-invariant long-range interactions (**Fig. S8C**), providing a reasonable template for weighting the contributions of different enhancers in the *MYC* locus across cell types.

To rank elements across the entire domain, we calculated the predicted impact score as described above in 400-bp windows tiled every 100-bp across chr8:127000000-131500000. DHS and H3K27ac were not always available for each of the 235 different samples — accordingly, we used both datasets where available, or calculated an alternative ranking using one or the other dataset (*e.g.*, DHS or H3K27ac normalized read count × normalized Hi-C signal). Given the varying patterns of DHS and H3K27ac signal around a regulatory element (DHS is strong at the center of the element while H3K27ac is depleted in the nucleosome-free region but strong just outside), we smoothed these scores at 2-kb resolution to better compare models generated from DHS or H3K27ac alone. To collapse neighboring windows with strong scores yet retain resolution for the strongest local maximum (*e.g.*, corresponding to the center of the regulatory element), we removed windows that had an overlapping window with a higher score. Finally, we assigned a rank to these remaining windows ("Enhancer Rank" column in **Table S1**), and focused on the top 10 elements in each cell type.

Analysis of enhancers known to regulate *MYC*

We curated a list of enhancers that have been shown to regulate *MYC* in their endogenous genomic contexts. (i) An enhancer implicated in *MYC* regulation in the context of colorectal cancer ("Myc-335") was identified based on an association rs6983267 and risk for colorectal cancer (*55*, *56*). Genetic knockout of this enhancer in mice leads to an ~40% reduction in Myc RNA expression in the colon, and confers resistance to intestinal tumorigenesis in an APC-/- background (*57*). (ii) An enhancer implicated in *MYC* regulation in the context of lung adenocarcinoma (LUAD) was identified based on a focal amplification of a noncoding region in multiple primary LUAD tumors (*22*). Genetic knockout of this enhancer in a LUAD cell line led to a ~30% reduction in *MYC* expression (*22*) and defects in cellular proliferation. (iii) An enhancer implicated in T-ALL was identified based on focal amplifications of a noncoding region ~1.47 Mb downstream of *MYC* (*58*). This enhancer contacts the *MYC* promoter as assayed by chromosome conformation capture, and a mouse knockout of this element leads to defects in thymocyte development and improved survival in the context of NOTCH1-induced leukemogenesis (*58*, *59*). (iv) An enhancer implicated in AML was identified on the basis of strong occupancy by Brg1 in a murine leukemia cell line, and is focally amplified in ~3% of human AMLs. This enhancer (E3) was shown to loop to the *MYC* promoter, and knockdown of Brg1 led to dramatic loss of *MYC* expression (*60*). We extracted coordinates from these previous studies and overlapped these coordinates with highly ranked enhancers in relevant cell types (**Fig. 3B**).

Analysis of GWAS variants near *MYC*

We downloaded a list of variants associated with human phenotypes from the GWAS Catalog at EBI (https://www.ebi.ac.uk/gwas/, accessed May 11, 2016). 121 associations are reported in chr8:127900000-131000000. We used HaploReg v4.1 (http://www.broadinstitute.org/mammals/haploreg/haploreg.php, accessed May 11, 2016) (*61*) to identify SNPs linked to the GWAS index SNP with r^2 >= 0.8 in the European population. The black boxes in **Fig. 3C** represent the span of all such SNPs for each variant, collapsed by phenotype to yield 66 unique associations between a human disease or trait and a genetic haplotype. We highlight three examples where these SNPs overlap elements predicted to regulate *MYC*. (i) A SNP linked to increased risk of Hodgkin's lymphoma, which has previously been noted to overlap with B-cell specific H3K27ac signals (*52*), overlaps an element that our heuristic predicts to be quantitatively among the most important for regulating *MYC* in B cell lymphoma cells (**Fig. 3D**). (ii) A SNP associated with bladder cancer risk is located in a conserved DHS element active in multiple gastrointestinal tissues, and thus may regulate *MYC* in bladder epithelial cells, for which chromatin data is not available (**Fig. 3D**). (iii) A SNP associated with height overlaps a glucocorticoid receptor motif in a conserved H3K27ac-marked element active only in chondrocytes (**Fig. 3D**). (DHS data from chondrocytes was not available). Although this SNP is located >1.9 Mb from *MYC*, it resides at the anchor of the long-range chromatin loop near e7 (**Fig. 2A**), suggesting that this SNP may affect height by altering the regulation of *MYC* in a chondrocyte-related cell type. Dozens of other predicted regulatory elements overlap disease-associated genetic variants near *MYC* and are listed in **Table S1**.

Software for data analysis and graphical plots

We used the following software for data analysis and graphical plots: R Bioconductor (version 3.0) (*62*), Gviz (version 1.10.11), gplots (version 2.17.0), GenomicRanges (version 1.18.4) (*63*), rtracklayer (version 1.26.3) (*64*), BEDTools (*65*), Integrative Genomics Viewer (version 2.3.26) (*66*), Pandas (version 0.12.0), Matplotlib (version 1.3.0), Biopython (version 1.61) (*67*), and SciPy (version 0.12.0).

Genome build

All coordinates are reported in human genome build hg19.

Supplementary Text

A generalizable method to discover and characterize gene regulatory elements

We set out to develop an approach to identify noncoding elements that regulate a given gene in its endogenous genomic context. A method to accomplish this would need to be able to (i) survey the regulatory function of many thousands of kilobases of genomic sequence, including regions not predicted to have regulatory function; (ii) sensitively identify and robustly quantify the effects of noncoding elements, and (iii) be generally applicable to study any gene of interest.

We designed our CRISPRi-based screening approach to address these goals. Our results in the *GATA1* and *MYC* loci demonstrate that this approach is scalable, sensitive, and specific. In the following sections we describe the conceptual and technical features that enable these characteristics and compare this method to similar approaches that use catalytically active Cas9 (*23*-*25*).

CRISPRi enables scalable functional characterization of gene regulatory elements. Because noncoding regulatory elements can be located far from their target genes and a gene might be controlled by multiple elements (*7*, *8*, *47*), a method to dissect the regulatory architecture of a given gene must be able to interrogate, through loss-of-function experiments, large regions of genomic sequence. To develop a scalable method, we exploited the programmable CRISPR system in the setting of a pooled screen to simultaneously interrogate the functions of many noncoding regions. In this method, we synthesize a library of sgRNAs targeting noncoding regions of interest; generate a lentiviral library containing each of these sgRNAs; and establish a population of cells in which each cell expresses doxycycline-inducible KRAB-dCas9 and a single sgRNA. The effects of each sgRNA can be identified by using high-throughput sequencing to characterize the representation of sgRNAs in the cell population before and after a phenotypic selection (*68*, *69*). This approach enables high-throughput interrogation of noncoding elements: in this study, we assay 1.29 Mb of sequence around *GATA1* and *MYC* in a single pooled experiment.

CRISPRi robustly identifies gene regulatory elements. A method for characterizing the regulatory network for a given gene needs to be able to robustly identify regulatory elements, even when their effects on gene expression are relatively small in magnitude. Several features of our approach help to provide high sensitivity and specificity for regulatory elements.

First, the pooled screening format provides numerous advantages that help to identify small effects. Specifically, pooled screens include contributions of many individual cells for each sgRNA; assess the functions of different sgRNAs in the same experimental context (in the same plate); and measure changes in sgRNA representation using count-based statistics.

Second, the use of the KRAB-dCas9 system enables independent assessments of the function of the same regulatory element with multiple adjacent sgRNAs. This property stems from the fact that KRAB-dCas9 appears to disrupt the functions of regulatory elements across distances on the order of hundreds of base-pairs (*12*), such that in the *MYC* and *GATA1* loci we observe regions where dozens of sgRNAs are consistently depleted (**Fig. 1B, 2A**). This is advantageous for quantifying the impact of an element because the efficacy of individual sgRNAs varies for reasons inherent to the CRISPR system, such as the effect of the targeting sequence on sgRNA transcription or stability (*68*). Thus, the degree to which an individual sgRNA affects gene expression reflects not only the importance of the disrupted element but also the potency of the sgRNA itself. To address this issue, we average the scores across multiple consecutive sgRNAs, providing a more robust estimate of the effect of an individual element. We note that this property appears to differ qualitatively from previous approaches using catalytically active Cas9 to perform mutagenesis of noncoding regions (*23*-*25*). Cas9-mediated mutagenesis relies on non-homologous end-joining to disrupt critical sequence motifs, and so – because the resulting indels are on the order of tens of bases or smaller – only the few sgRNAs very close to critical sequence motifs appear to disrupt the function of any given regulatory element (*23*-*25*). These properties may be important in determining the power of screens using each approach and may have different trade-offs for positive versus negative selection screens.

Supporting the specificity and sensitivity of this approach, we find that each of the elements identified by our CRISPRi screens (e-GATA1, e-HDAC6, and e1-e7), do in fact affect the expression of the intended gene, including effects on gene expression as small as 10%. We note that the sensitivity of this approach for even smaller effects might be accomplished by assaying more cells per sgRNA.

CRISPRi-based screening is general and can be applied to study other genes or phenotypes. A general method for identifying gene regulatory elements should be applicable to any gene of interest. While we looked for effects on survival and proliferation in K562 cells in order to characterize multiple gene loci in a single screen, we note that this CRISPRi-based approach could be applied to study an arbitrary gene of interest through fluorescence-based readouts of cells with a gene tagged in its endogenous locus with GFP (*23*). This strategy for mapping regulatory elements can also be applied in the context of other functional readouts, including other FACS-based assays (*24*, *70*) or drug or toxin resistance phenotypes (*10*, *69*).

Together, these properties provide a scalable, sensitive, and general method for mapping the functions of gene regulatory elements. This CRISPRi-based approach appears to have complementary properties to Cas9-mediated mutagenesis approaches (*23*-*25*): CRISPRi can robustly identify gene regulatory elements and provides non-mutagenic inhibition that is consistent across individual alleles and cells, while mutagenesis-based approaches appear to provide high resolution for identifying specific motifs. Further work will be required to determine how to best leverage these complementary features to dissect the networks of noncoding elements controlling gene expression. Finally, we note that in theory neither approach will be able to identify elements that act redundantly with other elements in a given locus, or elements that reside in repetitive genomic regions that cannot be uniquely targeted with CRISPR. Although we found several instances in which promoters repress neighboring genes, perhaps by a competition mechanism, it remains unclear whether CRISPRi can identify other types of repressive elements that are not promoters, and its utility in assaying intronic enhancers is unclear. Further technical advances will be required to characterize and explore the functions of these elements.

Essentiality of noncoding RNAs in the MYC locus.

Previous CRISPR screens have established that the protein coding genes expressed in the vicinity of *MYC* are not essential in K562 cells (**Fig. S1**). We further considered whether noncoding RNA genes in this region — including *PVT1*, *CCDC26*, and 5 microRNAs — are also essential and thus might explain the effects on cell proliferation conferred by the enhancers we discover in the *MYC* locus. In each case, we found that these noncoding RNAs either do not affect cell proliferation in K562 cells (PVT1 and CCDC26) or are not detectably expressed (microRNAs) and thus that e1-e7 likely control cell proliferation through regulation of *MYC*.

Two of the *MYC* enhancers we identified (e3 and e4) correspond to promoters that produce short alternative isoforms of the long noncoding RNA (lncRNA) PVT1 (**Fig. 2A**). Because PVT1 has previously been reported to affect cellular proliferation in *trans* based on siRNA-mediated knockdown experiments in mammary and ovarian cell lines (*71*, *72*), we investigated whether a *trans* function of the PVT1 transcript could be responsible for its promoters affecting cellular proliferation in K562 cells. We performed competition assays between K562 cells transfected with control siRNAs and cells transfected with siRNAs against PVT1 or, as positive controls, MYC or GATA1 (see Methods). Knockdown of MYC or GATA1 (27% or 52% reduction, respectively) led to a reduction in cellular proliferation relative to cells transfected with control siRNAs, as expected (**Fig. S1C,D**). In contrast, knockdown of PVT1 (66% reduction for the best siRNA) did not lead to detectable changes in proliferation (**Fig. S1C,D**). This indicates that reduction of the mature PVT1 lncRNA does not affect the proliferation of K562 cells.

In contrast, we found that CRISPRi targeting e3 (corresponding to a TSS of PVT1), which led to a ~77% reduction in PVT1 RNA levels (**Fig. S1E**), *did* affect cellular proliferation in competition assays (**Fig. 2C**). Thus, the proliferative defect observed upon inhibition of these elements in K562 cells appears to reflect their functions in the *cis* regulation of *MYC* rather than previously reported *trans* functions of the PVT1 RNA transcript itself. This is consistent with previous findings that gene promoters (including promoters of lncRNAs) can act as enhancers for neighboring genes (*73*, *74*). Indeed, we show that both e3 and e4 activate expression of a plasmid-based reporter gene (**Fig. S5B**, see Methods), indicating that these elements can act as enhancers. Further work will be required to investigate the possibility that other mechanisms associated with PVT1 transcription might also quantitatively contribute to controlling *MYC* expression in *cis*.

In addition to PVT1, the *MYC* region also contains the lncRNA CCDC26 (a pseudogene), which is expressed from a TSS 7.2 Kb distal to e5. Although e5 scored in our screen and affected *MYC* expression, we did not observe depletion of sgRNAs targeting the *CCDC26* TSS or promoter despite an abundance of sgRNAs in these regions (**Fig. S5B**). Thus, e5 and other enhancers likely affect cell proliferation through regulation of *MYC* rather than through regulation of *CCDC26*. We note that it is technically possible that depletion of *CCDC26* or *PVT1* contributes to affecting cell proliferation *in the context of MYC suppression*, but our data are inconsistent with them having strong effects on cell proliferation independent of changes in *MYC*.

The genetic region around also *MYC* harbors five putative miRNA genes previously described in several cancer cell lines (miR1204-1208). To determine if these miRNAs are expressed in K562s, we inspected ENCODE short RNA sequencing data (wgEncodeCshlShortRnaSeqK562CellShortAln.bam) and found that 0 reads (out of >29 million reads) overlap the RefSeq-annotated putative miRNAs in the region. Because regulation by miRNAs is thought to be highly dependent on miRNA abundance (*75*), miR1204-1208 do not likely have important functions in K562 cells.

Repressive elements in the *MYC* locus.

We identified 2 elements in the *MYC* locus (r1 and r2, **Fig. 2A**, **S5**) whose inhibition by CRISPRi led to *increased* proliferation of K562 cells in our screen, suggesting that these elements may act to repress *MYC* expression. Both of these elements have smaller absolute effect sizes in the screen data than the weakest detected enhancer (e5, 10% reduction in *MYC* expression), suggesting that these repressive elements may have even smaller quantitative effects on *MYC* expression. Interestingly, one of these elements corresponds to the promoter of a minor PVT1 isoform (**Fig. 2A**), consistent with a model wherein this promoter of *PVT1* competes with the *MYC* promoter for regulatory signals, similar to the phenomenon we observe for *GATA1* and *HDAC6*.

Conceptual framework for predicting enhancer function.

Our heuristic approach for comparing the relative activity of enhancers is based on a classic model in which an enhancer affects gene expression by recruiting transcription factors and activating gene expression upon physical contact ("looping") between the enhancer and a target promoter (*1*, *46*). In this model, the quantitative impact of an enhancer might depend on (i) its intrinsic activity (*i.e.*, the complement of transcription factors recruited to the element and their effects on a target promoter) and (ii) the frequency at which the enhancer physically contacts its target promoter in the nucleus. We note that this model does not represent all of the possible mechanisms by which regulatory elements might regulate their target genes (*1*), but does provide a simple framework with which to combine these two aspects of enhancer function.

To represent the intrinsic activity of an enhancer, we used quantitative measures of DHS and H3K27ac occupancy (see Methods) based on previous evidence that they correlate with various measures of activity. For example, DHS signal at regulatory elements in the genome correlates with transcription factor occupancy (*44*, *76*) and with the activity of those elements in plasmid-based reporter assays (*77*). H3K27ac occupancy correlates with expression of neighboring genes across cellular contexts (*78*, *79*) as well as with on-plasmid enhancer activity (*77*).

To represent the contact frequency between an enhancer and promoter, we used genome-wide measurements based on Hi-C (*80*) (see Methods), a method that requires physical contact and crosslinking in order to produce a signal linking two regions of genomic DNA. Physical contacts between enhancers and promoters correlate with gene activation (*1*, *6*, *46*, *47*), and in a few cases increasing the frequency of enhancer-promoter contact has been shown to activate gene expression (*48*, *49*).

These observations provide a conceptual foundation for this heuristic approach to comparing the relative impact of enhancers on gene expression. Further work will be necessary to determine whether this approach in fact reflects the mechanisms by which these enhancers regulate *MYC*. Regardless of the underlying mechanisms, this simple heuristic can distinguish elements that regulate *MYC* in K562 cells from those that do not and may be more broadly useful for connecting regulatory elements with their target genes.

Guidelines for design of additional CRISPRi screening libraries.

We sought to determine how to best design CRISPRi screening libraries using fewer sgRNAs per gene and thus enabling the interrogation of more genes. We analyzed our data by down-sampling the number of sgRNAs to every 2nd, 4th, 5th, or 10th sgRNA with each 20-sgRNA window. We found that, as expected, this reduces the reproducibility of estimates of the quantitative effects of elements and thus reduces power to detect elements with small effects (**Fig. S9A**).

An alternative strategy for designing smaller libraries is to focus on the subset of regions that are likely to score. All of the elements detected in our screen are centered on DHS sites (**Fig. S9B**) and every significantly depleted or enriched 20-sgRNA window is located within 1 kb of a DHS peak (the union of wgEncodeUwDnaseK562PkRep1.narrowPeak and wgEncodeUwDnaseK562PkRep2.narrowPeak). Designing a screen against only DHS sites could reduce the size of the library by approximately a factor of 5. However, it remains unclear whether there are regulatory elements in other loci that are not DHS sites.

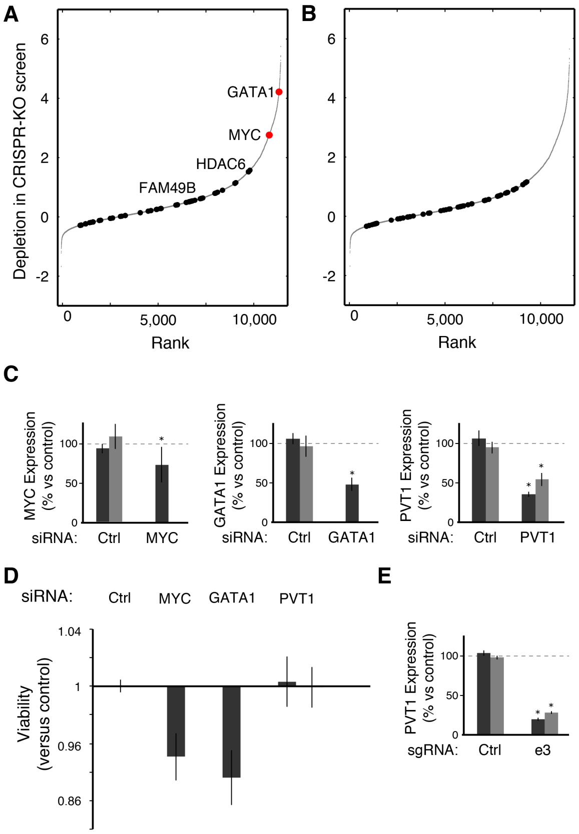

Fig. S1. ***GATA1* and *MYC* are encoded far from other genes that strongly affect proliferation in K562 cells.**

(A) Gray: Depletion ($-\log_2$ fold-change after 14 population doublings) in a previous genome-wide CRISPR knockout screen of all genes expressed in K562 cells (*26*). Higher scores denote stronger effect on proliferation. Black: genes within 500 Kb or in the same topological domain as *MYC* or *GATA1* (highlighted in red).
(B) Same for the three tiled negative-control regions.
(C) Knockdown efficiency for siRNAs targeting MYC, GATA1, and PVT1, as assayed by qPCR compared to siRNAs without an RNA target (Ctrl). Gray bars: two different siRNAs for Ctrl and PVT1. Error bars: 95% confidence intervals (CI) for the mean of four independent transfections. *: $p < 0.05$ in T-test versus negative controls.
(D) Relative viability of cells in a competitive growth assay (gamma). GFP-expressing cells were transfected with siRNAs against GATA1, MYC, PVT1, or siRNAs without a genomic target (Ctrl) and were mixed with RFP-expressing cells transfected with a Ctrl siRNA and grown for four days before counting. Error bars: 95% confidence intervals (CI) for the mean of 4 independent transfections. We tested two different sgRNAs for PVT1. *: $p < 0.05$ in T-test versus negative controls.
(E) qPCR for PVT1 RNA in cells expressing sgRNAs targeting a TSS of PVT1 (e3) or sgRNAs without a genomic target (Ctrl). KRAB-dCas9 expression was activated with doxycycline for 24 hours before measurement. Gray bars: two different sgRNAs per target. Error bars: 95% confidence intervals (CI) for the mean of 3 independent infections. *: $p < 0.05$ in T-test versus negative controls.

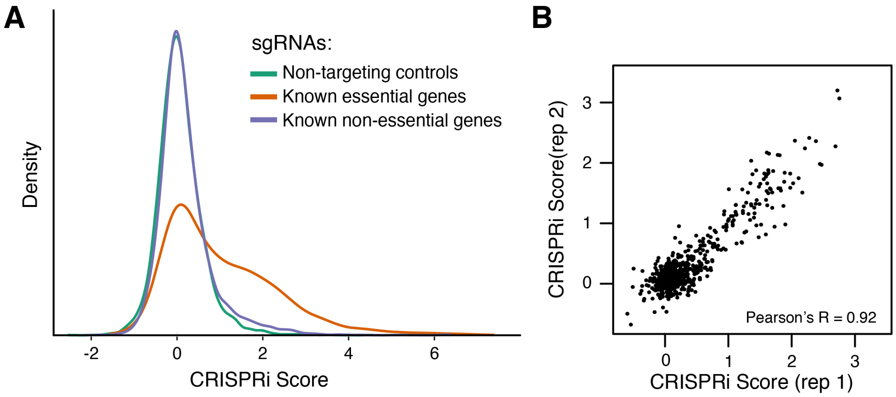

Fig. S2. CRISPRi screen reproducibly depletes sgRNAs targeting promoters of essential genes.

(A) Distributions of CRISPRi scores for sgRNAs targeting the promoters of genes previously identified as essential or non-essential based on a genome-wide CRISPR knockout screen (*26*) and for sgRNAs with no genomic target (control sequences). A higher CRISPRi score indicates stronger depletion over the course of the screen.
(B) Average CRISPRi scores for 600 protein-coding gene promoters in replicate screens.

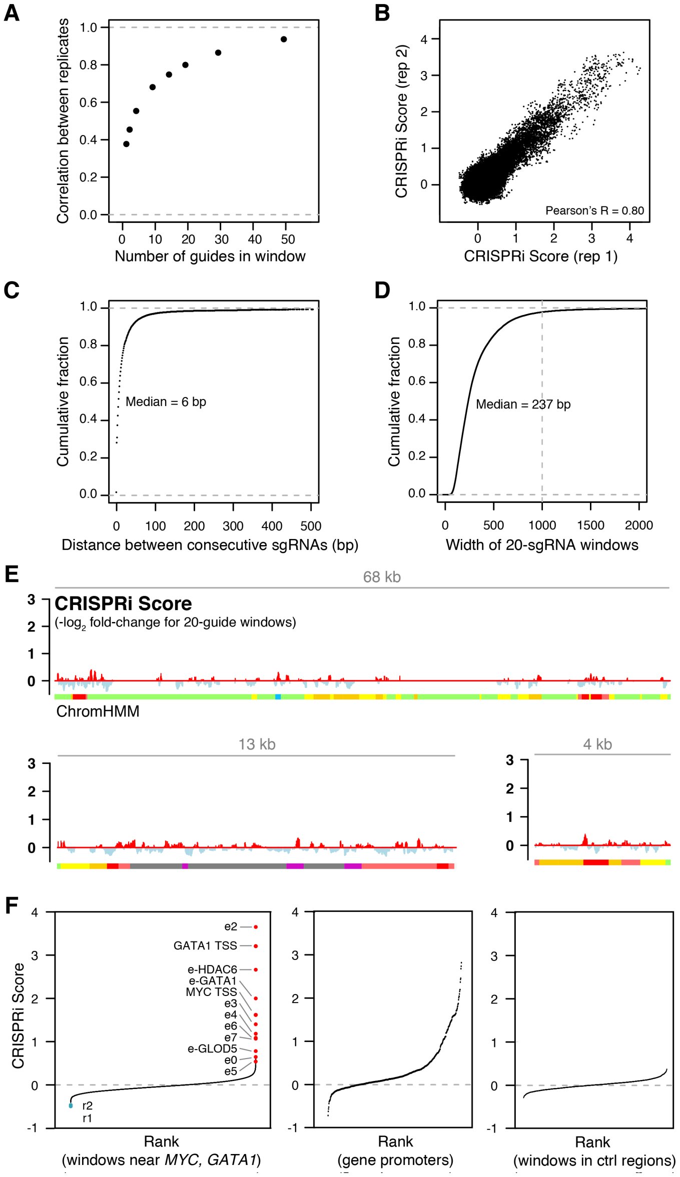

Fig. S3. Sliding window approach for analyzing CRISPRi screens.

(A) Pearson correlation between the two replicate screens for CRISPRi scores averaged across windows of different sizes (2, 3, 5, 10, 15, 20, 30, or 50 consecutive sgRNAs).
(B) CRISPRi scores for all windows of 20 consecutive guides in the replicate screens.
(C) Cumulative density plot of the distance between consecutive sgRNAs. Distribution extends beyond the *x*-axis limits.
(D) Cumulative density plot for the span of 20-sgRNA windows. Windows spanning greater than 1 kb were not considered. Distribution extends beyond the *x*-axis limits.
(E) CRISPRi scores in 20-sgRNA windows for three negative-control regions that are located far from known essential genes (see Methods). These regions show a lack of strong signal as compared with the *GATA1* and *MYC* loci and were used to calculate an empirical false discovery rate for the CRISPRi score.
(F) Gray: CRISPRi score in 20-sgRNA windows for tiled *MYC* and *GATA1* regions (left, ~60,000 windows), the TSSs of protein coding genes from across a range of essentiality (middle, ~600 genes), or tiling regions far from any essential gene (right, ~5,000 windows). Red dots: Most strongly depleted window within identified enhancers and TSSs (other windows nearby, which are also often strongly depleted, are not shown for visual clarity). Blue: Most strongly enriched window within putative repressive elements.

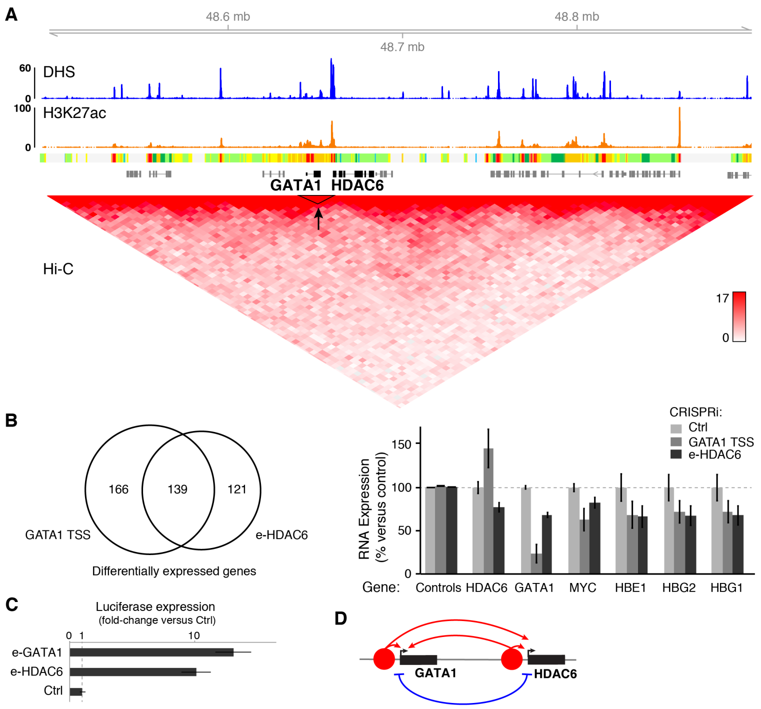

Fig. S4. Characterization of enhancers at the *GATA1* locus.

(A) Chromatin state and chromosome conformation in the ~400-Kb topological domain containing *GATA1* and *HDAC6*. K562 DHS, ChIP-Seq data, and chromatin state classifications (ChromHMM) are from ENCODE (*4*) (see Methods). Contact frequency matrix is derived from *in situ* Hi-C maps at 5-kb resolution in K562 cells (KL-normalized observed matrix) (*6*). Black triangle and arrow mark the region of interactions between enhancers (e-GATA1 and e-HDAC6) and the promoters of GATA1 and HDAC6.

(B) Effects of inhibiting *GATA1* TSS or e-HDAC6 on gene expression of downstream GATA1 target genes. Venn diagram represents differentially expressed genes from RNA sequencing of stable lines expressing the listed sgRNA relative to cells containing negative control sgRNAs (Ctrl). Hypergeometric *p*-value of overlap $<10^{-163}$. Bar plot shows that known target genes of the GATA1 transcription factor (MYC, HBE1, HBG1, and HBG2) (*81*-*83*) are differentially expressed upon inhibition of e-HDAC6. KRAB-dCas9 expression was activated for 24 hours before measurement. Error bars: 95% CI for the mean of 2 sgRNAs with 3 independently derived stable lines each. Controls: all other expressed genes.

(C) Expression of firefly luciferase from plasmids containing each enhancer located 2 kb upstream of a *MYC* promoter fragment. Data is normalized to a random sequence of similar size (Ctrl) and to the internal *Renilla* luciferase control (see Methods). Error bars: 95% CI for the mean of 3 independent transfections.

(D) Regulatory connections in the *GATA1*/*HDAC6* locus: two enhancers (red) regulate both genes, and the promoters appear to repress one another (blue), perhaps by competing for activating signals from the enhancers.

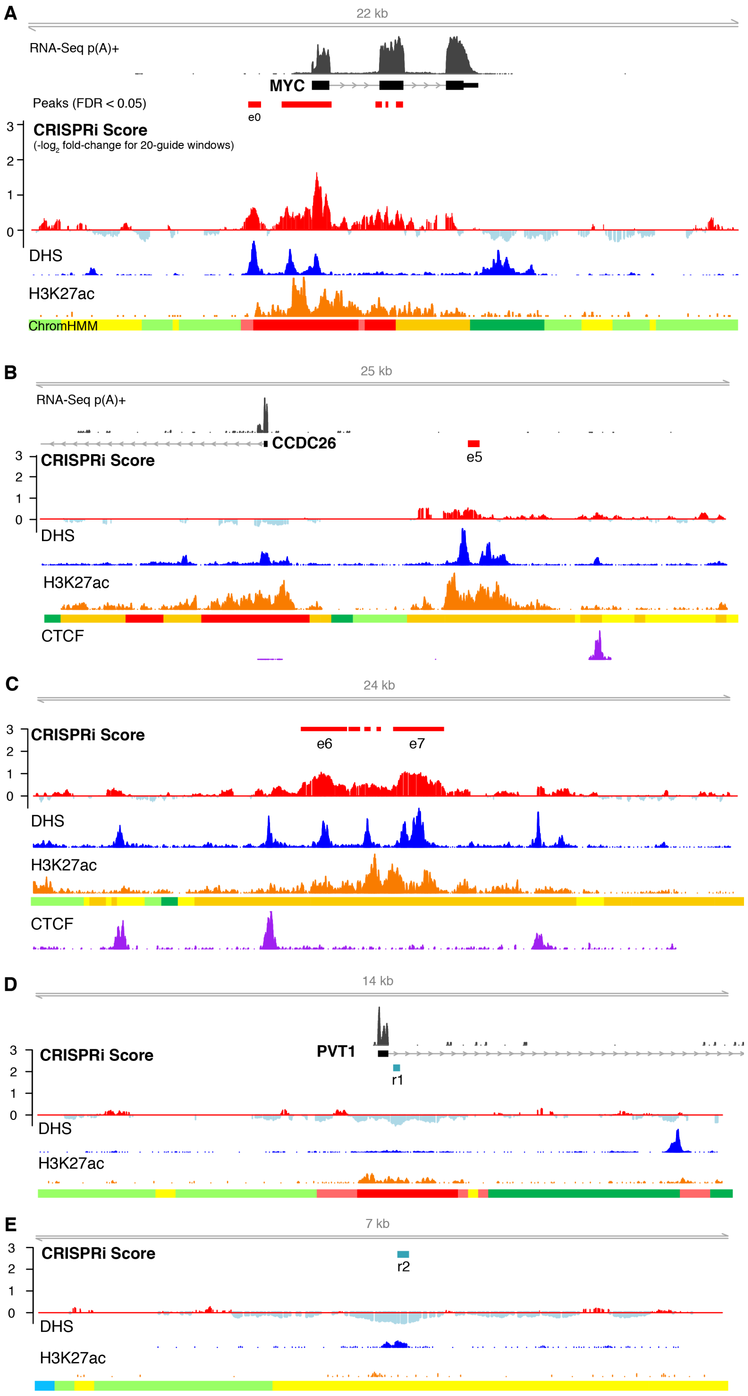

Fig. S5. Regulatory elements at *MYC* and downstream enhancers.

(A) CRISPRi screen results in *MYC* gene locus, showing significant peaks at the *MYC* TSS, at several locations in the gene body, and at a known promoter-proximal regulatory element (e0) (*21*). K562 DHS, RNA-Seq, ChIP-Seq data, and chromatin state classifications (ChromHMM) are from ENCODE (*4*).

(B) Expanded region around e5 and CCDC26 and **(C)** e6/e7 showing strong CTCF occupancy at DHS sites close to the elements. Each CTCF peak has a motif oriented in the reverse direction (toward *MYC*, not pictured, see Methods). Note that the promoter of CCDC26 does not score as essential, indicating that its expression is not responsible for the proliferative defects observed upon inhibiting e5 or other enhancers.

(D) Expanded region around the putative repressive elements r1 and **(E)** r2. r1 corresponds to the promoter of an alternative isoform of PVT1.

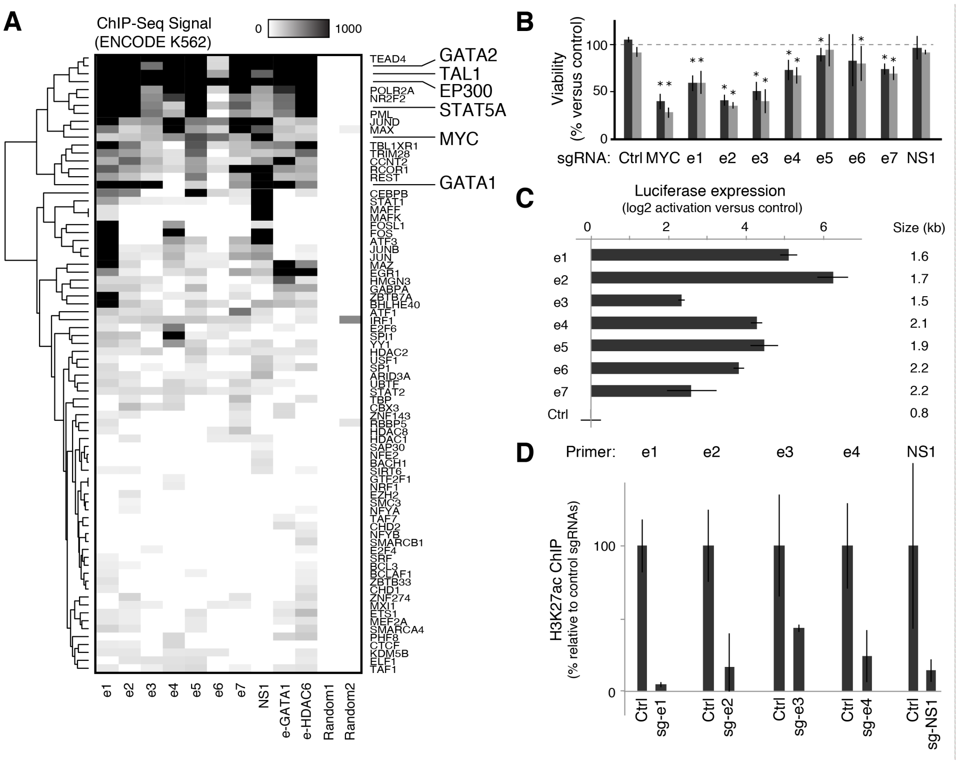

Fig. S6. Characterization of enhancers at the *MYC* locus.

(A) *GATA1* and *MYC* enhancers bind many activating transcription factors. Transcription factor binding in a 1-kb window centered on each enhancer are shown with their ChIP-Seq signal reported by ENCODE (*4*), which assigns scores to peaks by multiplying the ChIP-seq signal values by a normalization factor calculated as the ratio of the maximum score value (1000) to the ChIP-seq signal value at one standard deviation from the mean, with values exceeding 1000 capped at 1000. For comparison, two random sites near *MYC* are shown.

(B) Relative viability of cells in a competitive growth assay. Cells expressing the indicated sgRNAs were competed against K562 cells expressing GFP or RFP and grown in doxycycline for 7 days before counting. Gray bars: two different sgRNAs per target. Error bars: 95% CI for the mean of 6 total replicate competition assays using cells from 3 independent infections. *: $p < 0.05$ in T-test versus negative controls.

(C) Each *MYC* enhancer can activate a reporter gene driven by a *MYC* promoter fragment in a plasmid-based luciferase assay. The size of each enhancer sequence is reported on the right. Ctrl: negative control sequence corresponding to a bacterial kanamycin resistance gene. Error bars: 95% CI for the mean based on three replicate transfections.

(D) To determine if sgRNAs targeting NS1 successfully affected chromatin state, we performed ChIP for H3K27ac in cells expressing individual sgRNAs targeting e1, e2, e3, e4, or NS1, as well as two non-targeting control sgRNAs (see Methods). We measured ChIP enrichment by qPCR for 5 positive control loci, 3 negative control loci, and the locus targeted by the sgRNA (see Methods). Bars represent enrichment of the indicated locus normalized to the non-targeting control sgRNAs. Error bars: 95% CI for the mean for 5 (Ctrl) or 3 (others) biological replicates.

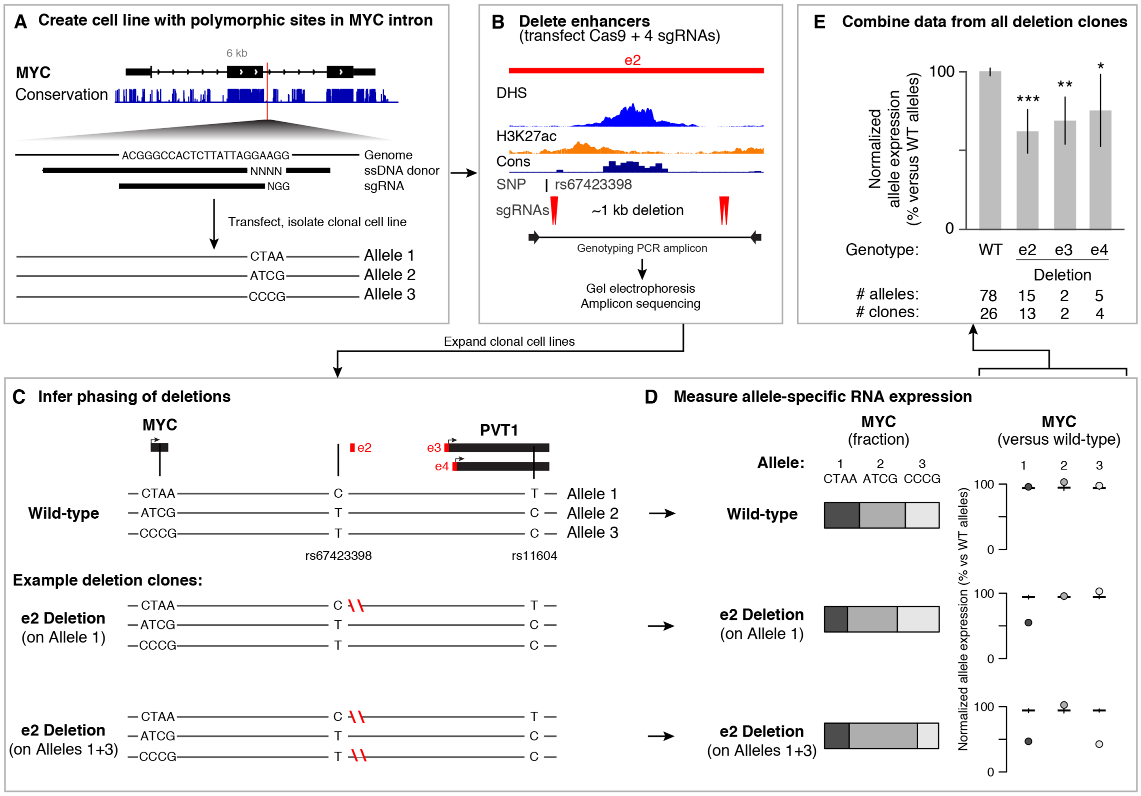

Fig. S7. Genetic deletions of enhancers in the *MYC* locus.

(A) Strategy for generating a cell line containing polymorphic sites on each allele of *MYC*. We used CRISPR/Cas9 to knock in a random 4-mer sequence into an intronic site in the *MYC* locus that was not conserved across mammals (red line). We co-transfected a plasmid expressing Cas9, a ssDNA oligo donor, and an sgRNA, picked clonal cell lines, genotyped by amplicon sequencing, and isolated a clone with three unique alleles.
(B) Strategy for deleting enhancers, showing e2 as an example. To delete each enhancer, we designed 4 sgRNAs flanking the DHS peak in the center of each element, two on each side. We co-transfected these 4 sgRNAs and isolated clones containing deletions on 1 or 2 of the 3 alleles. The rs67423398 SNP was contained in the genotyping PCR amplicon and was used to determine which allele of e2 was deleted.
(C) Overview of sites relevant to enhancer deletions in the *MYC* locus, including inferred phasing of polymorphic sites. Bottom: Genotypes for example deletion clones.
(D) Allele-specific RNA measurements for representative clones. For each clone, we determined the fraction of RNA molecules carrying each of the *MYC* alleles using ddPCR (bar plots). We calculated a fold-change for each allele in deletions versus controls and normalized this to the highest of these three values within each clone (see Methods). This yielded the "normalized allele expression" (right). Dots: values for one clone. Horizontal bars: mean with 95% confidence interval for 26 wild-type clones.
(E) Deletions of e2, e3, and e4 led to a 30-40% decrease in the expression of *MYC* on the corresponding allele compared to wild-type alleles in the same cells. We compared normalized allele expression values between wild-type and deletion alleles using a Wilcoxon rank-sum test. *: $P < 0.05$. **: $P < 0.01$. ***: $P < 10^{-4}$.

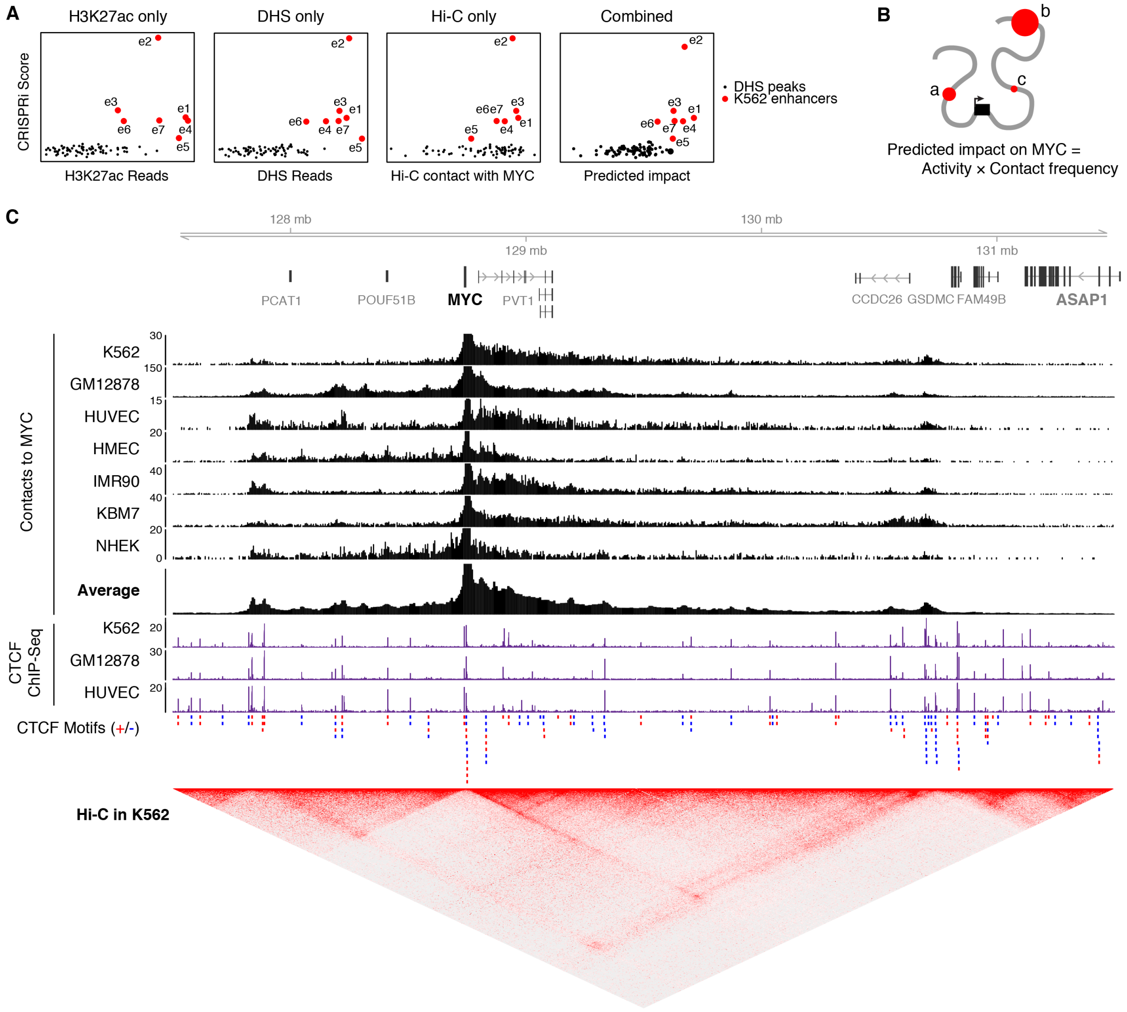

Fig. S8. Heuristic model for predicting enhancer function in the *MYC* locus.

(A) Comparison of models using H3K27ac only, DHS only, Hi-C only, or a combination of all three (Predicted Impact, same as Fig. 2E). This ranking is applied to 93 elements selected based on DHS and H3K27ac signal (see Methods), and thus provides an optimistic estimate of the power of each individual source of information for predicting MYC enhancers.

(B) Heuristic framework for predicting the relative impact of regulatory elements on *MYC* expression. Impact depends on activity (estimated by quantitative H3K27ac and DHS signal, represented by size of red dot) and the frequency with which it contacts the *MYC* promoter (estimated based on Hi-C, represented by distance from gene). For the three example enhancers, their relative impact would be a = b > c.

(C) Comparison of Hi-C and CTCF ChIP-Seq signal in the *MYC* locus across cell types. Contact frequency with the *MYC* promoter is derived from *in situ* Hi-C maps at 5-kb resolution across 7 cell types (KL-normalized observed matrix) (*6*). *Y*-axis differs between cell types according to the depth of sequencing. The average contact profile used in our enhancer ranking calculations across cell types was created by averaging the normalized contact frequencies from these 7 cell types. CTCF motifs are colored according to their orientation: red = positive strand, blue = negative strand (see Methods).

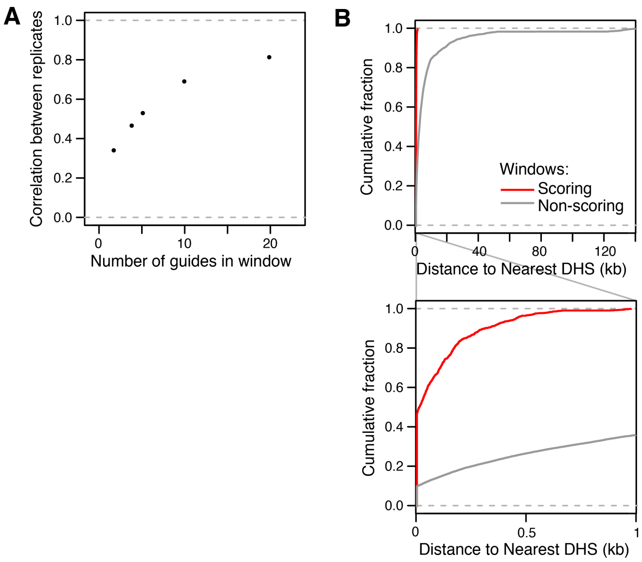

Fig. S9. Design of new CRISPRi libraries

(A) Pearson correlation between the two replicate screens for CRISPRi scores from windows of different sizes – 2, 4, 5, 10, 20 sgRNAs – downsampled by taking every 10th, 5th, 4th, 2nd, or every sgRNA, respectively. Reducing the density of coverage reduces reproducibility.

(B) Cumulative density plot of the distance between 20-sgRNA windows and the nearest DHS peak, with the first kb highlighted below. All significantly enriched or depleted windows (Scoring) are less than 1 kb from a DHS peak, compared to <35% of all other windows (Non-scoring).

Table S1 (separate file)

Trait-associated polymorphisms in predicted *MYC* enhancers across cell types.
Genetic variants linked to human traits overlap regulatory elements predicted to regulate *MYC*.

Table S2 (separate file)

CRISPRi sgRNA library sequences and screening data.
Sequences, annotations, CRISPRi scores, and raw counts for sgRNA library.

Table S3 (separate file)

Sequences of primers, oligos, sgRNAs, siRNAs, and ddPCR probes.
(A) Primer sequences for RT-qPCR, ChIP-qPCR, and ddPCR.
(B) sgRNA sequences for single sgRNA CRISPRi, MYC tag knock-in, and enhancer deletion.
(C) Catalog numbers (GE Dharmacon) for siRNAs for *PVT1*, *MYC,* and *GATA1* knockdown.
(D) Cloning and sequencing primers for pooled sgRNA library.
(E) Primers for cloning enhancers for luciferase assays, generating the MYC-Tag cell line, and genotyping enhancer deletion clones.
(F) ddPCR probes for measuring allele-specific expression of *MYC* and *PVT1.*